Sentiment Analysis for Thai Natural Language Processing

Kritsada Sriphaew
Japan Society for the Promotion of Science
kong@lr.pi.titech.ac.jp

Hiroya Takamura and Manabu Okumura
Precision and Intelligent Laboratory
Tokyo Institute of Technology
takamura@pi.titech.ac.jp, oku@pi.titech.ac.jp

What other people think is an important piece of information for the most of customers during the decision-making process. Although a large amount of opinions are available online in the blogs, reviews and dialogues, it is not easy to read all of them and make a decision. Therefore, the sentiment analysis is an emerging research in natural language processing field that can help to automatically grasp the opinions on a particular product or topic that are available in texts. Three main tasks in sentiment analysis are (1) subjective/objective identification, (2) sentiment classification and (3) feature/aspect-based sentiment analysis. Several intensive works for corpus collection, labeling and developing the basic tools for natural language processing are also needed. Although most of the potential works on sentiment analysis focus on analyzing the opinions in English text, it is possible to apply them to Thai textual data by using some language-specific information, labeled data on Thai corpus and machine translation services. Sentiment analysis are not scoped within only business area, but it can be extended to grasp the opinions of any kinds of topic such as news, events or politic.

1. Introduction

Computers can work well with numerical computing, but can they grasp the feeling? Nowadays, personal opinions can be easily expressed via several forms of online texts such as blogs, reviews and dialogues. How to make the computer understand the opinionated text is a challenge research in natural language processing (NLP) field known as sentiment analysis or opinion mining [1].

With the fast-growing mountain of the opinionated texts available online, it is interesting to retrieve the plausible opinions of the customers on a specified product or particular topic. What other people think is an important piece of information for the most of customers during the decision-making process. Several customers make the decision after they read the product reviews. Some voters may change their minds after they are convinced by the opinions of the other voters.

For a popular product, the number of reviews can be in hundreds or even thousands. This makes it difficult for both a customer to read all of them and manufacturer of the product to keep track and manage customer feedback. Therefore, automatic extraction and summarization a large amount of opinions are required to facilitate the users. The objectives of the system are to detect the opinionated text from a large amount of textual data, determine whether such opinions are positive or negative and make a summary.

There are three main tasks in sentiment analysis: (1) subjective/objective identification that tries to identify whether a given text is objective or subjective, (2) sentiment classification that aims to classify whether a given text has positive or negative opinions, and (3) feature/aspect-based sentiment analysis which focuses to determine the opinions on each feature or aspect of the products. These tasks have been intensely explored in previous works to find the high accurate methods, and they were evaluated by the labeled corpus on some specific domains such as movie reviews, product reviews and political polls.

Although most of the potential works have been performed on English corpus, a resource-rich language, it is possible to do a sentiment analysis on some other languages by applying the existing methods with some language-specific information, labeled of the target languages and machine translation services.

2. Sentiment Analysis for Thai

The main task of the sentiment analysis concerns with the classification. For subjective/objective identification, the simplest way to classify whether the given text is subjective or objective is to use the terms or structure of text as cues for identification. For example, a sentence that contains the term which express the feeling such as "I think that" or "I feel that", is usually subjective sentence, or texts that are under the topic of "review" or "comment" can be assumed as an opinionated texts. To apply this for Thai, we can directly define such lexical cues in order to detect the subjective sentences from the objective ones, but the pre-process of word segmentation and sentence boundary detection must be applied beforehand.

For sentiment classification, several techniques have been developed to find out the semantic orientation of the opinion. The orientation can be classified into three classes, i.e., positive, negative and neutral. Sentiment classification can be performed in different levels of granularity of text, i.e., word, phrase, sentence, paragraph or document. Most of the techniques are based on machine learning approach where the labeled data is provided for learning the classification model. This is a main obstacle to the resource-scarce language such as Thai since such labeled data is not available and it consumes considerable time and large human efforts if we want to construct one. However, some techniques for cross-lingual analysis [2] can make it feasible by using machine translation services as tools to translate the English labeled corpus to the other target languages and applying learning technique for cross-learning on the corpus in both languages for the robustness of the model.

For feature/aspect-based analysis, the features or aspects of the entities or topics are extracted with their underlying opinions. A feature or aspect can be an attribute, component or a function of an entity. For example, the picture quality, size and weight can be the features of a camera. To implement this task, mining technique is applied to extract the features or aspects of the entities by finding the noun phrases that are usually occurred with the terms that express the opinion [3].

3. Discussion

Although it is possible to apply the existing methods for sentiment analysis on English corpus to Thai textual data, the performance is still far from satisfactory because of the language gap and language-specific characteristic. Besides, Thai labeled data for sentiment analysis is required as a benchmark for evaluation. To establish this research for Thai NLP community, several intensive works are also needed such as corpus collection, labeling and developing the basic tools for NLP (e.g., word segmentation, part of speech tagger, sentence boundary detector). Moreover, one important related field that can help us to fasten the development process by transferring the technology implemented in the resource-rich language to the resource-scarce language is a machine translation. However, the imperfection of the translation is still a main factor that affects the system accuracy.

4. Conclusions

The word of mouth is a critical matter that plays an important role for the business. The positive opinions can help to attract more customers to buy the products while the negative opinions can help the manufacturers to improve the products. To apply sentiment analysis, we gain not only what other people thinks on a particular topic or product, but also some new ideas for improving the products or services by spending only smaller budget and time than making an actual survey. This matches with most of the Thai business where budget and time are limited. Moreover, sentiment analysis are not scoped within only business area, but it can be extended to grasp the opinions of any kinds of topic such as news, events or politic for the benefits of understanding the society or making any change to the community.

References

[1] B. Pang, L. Lee, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval 2(1-2), 2008, pp. 1–135.

[2] X. Wan, Co-training for cross-lingual sentiment classification, ACL, 2009, pp. 235-243.

[3] M. Hu, B. Liu, Mining and summarizing customer reviews, KDD, 2004.